

The performance of credit risk evaluation models Empirical aspects and metrics

Laura POPESCU

Academy of Economic Studies, Bucharest, Romania

laura.popescu@brd.ro

Abstract: *Credit risk evaluation has become a topic of great importance in the context of the current economic crisis which has brought the financial institutions on the point of dealing with an increased risk of granting credits to defaulted clients. Consequently, the need to use efficient evaluation methods is an increased one. Researchers have designed and studied a large variety of analysis models based on expertise, mathematical algorithms or artificial intelligence techniques. All the results have concluded that one cannot determine a unique evaluation tool that is able to win the competition with the other methods under a large variety of circumstances. In order to obtain a result that can be associated with a high confidence level it is compulsory to take into consideration several characteristics. The current paper aims to present several techniques that can be used in order to evaluate the performance of credit risk evaluation models. Some important empirical aspects for performance estimation are described: the data sets' analysis, discrimination and stability. At the same time the article introduces some performance metrics: the execution time, overloading, efficiency, acceleration and costs. The importance of these tools is emphasized in the process of correctly estimating the advantages and disadvantages of each available methodology.*

Keywords: *performance, risk, evaluation, metrics, credits*

1. Introduction

The risk of granting credits has a very high importance when analyzing the activity of credit institutions as clients' selection process has a great influence on their financial results. Romanian banks have developed more and more complicated models that aim to catch the deviations of the models from reality. But patching a model to increase its complexity may not be an optimal way of modeling. Any economic system is extremely complicated but, to a large degree, this is due to the enormous number of degrees of freedom.

The accuracy plays an increasingly important role in performance measurement processes. During economic crisis, the behavior of credit institutions is different than in a steady or economic growth period. Therefore, it is very important to use all the available means to improve the quality of the tool that decides whether a credit is granted or not.

The easiest way to determine the performance of a credit risk evaluation model is to compute its success rate as a ratio between the correctly classified cases and the total number of entities. But in order to improve the quality of this estimator, one has to take into account the quality of the data set, the power of the algorithm to correctly evaluate the data outside the test sample, and the optimization of the process (the ability to obtain similar results on different samples).

Another aspect that needs to be taken into account refers to the performance metrics that are able to quantify the speed of the evaluation application (the faster the result is obtained, the better), its acceleration and the associated costs.

2. Empirical aspects in performance analysis

There are several empirical aspects that have to be taken into account when measuring the performance of the process of credit risk evaluation:

- The data set analysis
- Discrimination
- Stability

Data sets

In the process of measuring the risk associated to a credit applicant, it is necessary to undergo a stage that should be developed before applying complex computational algorithms or using expert systems or artificial intelligence based applications. It is recommended to perform an analysis of the model's quality. An important first step consists of analyzing the quality of the data set used for applying the model. The obtained results (output) and their accuracy depend on the quality of the dataset (input). Some important aspects have to be taken into account as they prove to be crucial:

- The usage of a large dataset.
- The data set contains a large number of clients (commercial clients or consumers) that have defaulted on their credits. This is important as it increases the ability to estimate whether the model has correctly classified the entities.
- The existence of a limited number of missing values for the characteristics taken into consideration for the components of the selected dataset.

In order to test the performance of a specific credit risk evaluation tool it is necessary to have a large dataset that increases the prediction power of the model for as many cases as possible. The database has to benefit from the necessary hardware resources in order to be able to deposit a large number of records (a record represents a set of characteristics for a certain client). Each record usually needs a considerable memory space as the characteristics/fields are qualitative, as well as quantitative.

The historical datasets also have an increased importance. Consequently, it is recommended not to completely erase from the system the data regarding the past evolutions of certain clients. Though periodic database cleaning actions are organized, all the information is backed-up, organized into archives, as it will prove its utility in the process of estimating the performance of the model. The method is applied for the historical data and it is determined the number of situations in which the model leads to an accurate estimation (the case in which a good client has been classified as Accepted) and the number of unsuccessful classifications (an application has been "Accepted", but the loan has finally defaulted).

The **Rate of Success** is quantified through the usage of the **RS** metric. The **RS** metric is computed as:

$$RS = TCC/TEC$$

where:

- **RS** represents the Rate of Success for the estimation tool
- **TCC** represents the Total Number of Correctly Classified Cases
- **TEC** represents the Total Number of Evaluated Cases

If the rate of success has a higher value, the tool is more performant. The financial institution has a certain risk aversion that influences the limit level of the success rate. A higher value of this cut-off level indicates a higher risk aversion.

Testing the classification tools on the historical data refers to the capacity of these methods to correctly classify the data that has already been “viewed”, meaning that the analysts already know the behaviour of the client during the whole life of the granted loan. The next section of the paper introduces an empirical aspect that refers to the capacity of correctly classifying the “unviewed” data, meaning the data located outside the selected dataset (the rest of the statistic population).

Discrimination

This concept refers to the ability of a credit risk evaluation model to correctly classify the “unviewed” data (those entities that are part of the total population, but have not been selected in the sample). It practically consists of the ability of drawing a clear separation line between the potential classes (successful cases versus defaulted cases). The objective is to obtain an error coefficient that does not depend on the selected set of records. This coefficient of default will lead to different values depending on whether the chosen method is applied on a training set versus a testing set. There are four procedures used to determine this **Coefficient of Default**:

1. “hold-out estimate error”

The selected items are divided into two sets: the training set that is used for training (minimizing the error inside the sample) and the testing set which is used in order to test the performance of the calculus (determining the external error for the sample). This estimation has the tendency to inflate the real value of the error. The dimension of the test sample is essential in order to diminish this problem.

In numerical analysis, the speed at which a convergent sequence approaches its limit is called the rate of convergence. This concept is of practical importance if we deal with a sequence of successive approximations for an iterative method, as then typically fewer iterations are needed to yield a useful approximation if the rate of convergence is higher.

The rate of convergence usually varies depending on the chosen credit scoring model. It is advisable to define as a logit the degree to which a certain model fits a certain situation. The main argument is that this type of function is more appropriate than other functions such as the probit or the k-nearest neighbour and will depend on the highest rate of convergence obtained from the available data.

2. “n-fold cross-validation method”

This technique consists of dividing the sample into n data sets. The training is performed on n-1 out of the n data sets and the testing is executed on the remaining sample. The same procedure is applied for n times as every sample has to become in its turn the testing subset.

3. “jackknife”

This procedure is used in order to reduce the error coefficient deviation. The basic idea of the procedure consists in a systematic recalculation of the statistic metric, but each time a

unit is deleted from the test sample. From the new data set a deviation approximation is determined. At the same time an estimation of the metric's variance is computed.

4. *the "bootstrap" technique*

This type of procedure offers variance and standard deviation estimations for the error coefficient that do not depend on any parameters due to the new observation samples generated through this technique. These results present a lower variance than those obtained through other methods.

Golfarelli, Maio and Maltoni (1997) have introduced the concept of error-reject tradeoff [1]. This concept is associated to those entities that have been classified with a very low confidence level. The global performance of the metric is a low one if the entity that needs to be classified is located very close to the border between the classes in which it can be placed. On a graphic representation this type of entities are placed in the area where the two classes intersect. Consequently, instead of deciding in which of the the categories the analysed case should be classified, the entity is rejected and the error coefficient is reduced. The higher the rejection rate, the lower is the error coefficient. A good technique to define a rejection rate consists of taking into consideration factors such as the trade-off and the costs associated with the rejection cases and incorrect decisions.

Henderson [6] emphasizes that the bootstrap technique is useful for analysing small expensive-to-collect data sets where prior information is sparse, distributional assumptions are unclear, and where further data may be difficult to acquire.

The last two methods (bootstrap and jackknife) estimate the variability of a statistic metric when the sample for which it is computed modifies. The jackknife method is a less general technique than the bootstrap method and has a different approach for the same variation. The jackknife method is easier to apply for the more complex samples and in many cases both methods will lead to similar results. If they are applied in order to compute the standard deviation of the metric, the bootstrap technique leads to slightly different results in repeated applications for the same data set, while the jackknife method obtains identical results for every iteration (under the hypothesis that the excluded subsets are the same)

Stability

This concept refers to optimization. When a user develops a credit risk evaluation model that has certain parameters or is independent, the goal is to obtain optimal results that are not influenced by the selected sample.

When it comes to parametric models, this characteristic refers to the selection of the most significant classes and factors that influence the loan granting process. Consequently, it is very important to establish the appropriate weights for each of the selected factors. The ability to obtain an unique optimal result and not only a local optimal solution is a very important aspect that requires additional testing. For the parametric and nonparametric models, the stability leads to a careful review of the treatment and of the impact over the model.

The least complicated methods consist of eliminating the extreme values from the selected sample, as they do not belong to the homogenous part of the sample: they are placed at a distance which is twice or three times greater than the standard deviation (**fig. 1**).

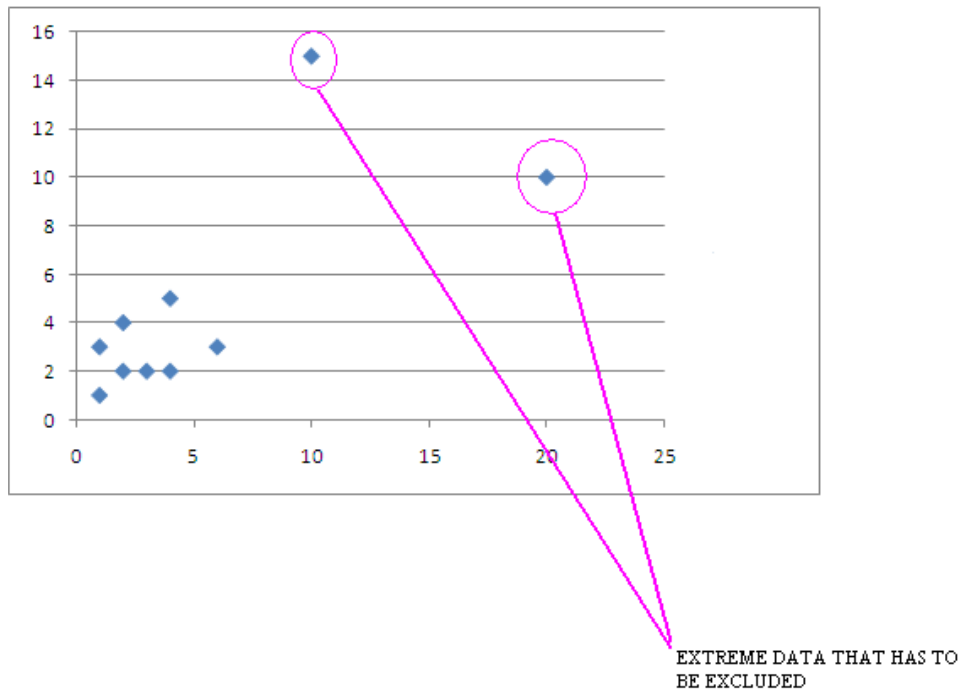


Fig. 1. Excluding the extreme values from a homogenous sample

Another technique which is also called “winzorization” consists of reintroducing the extreme values at a defined level which is previously mentioned. More complex mathematical methods such as introducing “wavelets” have also been developed, but they are usually applied for large time series, though they can be adapted for the “cross-sectional” cases.

As Struzik states [3], working without an a priori assumed model is typical for the modern approach in economics. Instead, the model is determined and reviewed step by step from the data sets. The data is analysed in terms of very generic analysis methods like the wavelet decomposition. The wavelet transform components are then analysed and simple or complex models can be created to fit the decomposition components. Consequently, the scaling of models or distributions can be tested and a hypothesis drawn.

3. Performance metrics

In order to evaluate the performance of credit risk evaluation applications, several metrics can be applied.

The execution time represents an important evaluation criteria. There are two ways of quantifying it: the serial execution time (TS) and the parallel execution time (Tp). The serial execution time is the time measured from the beginning to the end of the algorithm’s execution on a computer where the tasks are performed sequentially. The parallel execution time is computed as the duration of the algorithm in the case when the subtasks are parallel.

Wilhelm & co. [4] define the *Worst Case Execution Time (WCET)* of a computational task is the maximum length of time the task could take to execute on a specific hardware platform. Knowing worst-case execution times is of prime importance for the schedulability analysis of hard real-time systems.

The overloading is a metric computed as the difference between the total execution time of all the processes and the necessary time for the fastest sequential algorithm :

$$TO = p \cdot Tp - TS$$

where :

- p represents the number of processors
- $p \cdot Tp$ represents the time consumed by the p processors in order to complete the algorithm, to communicate between them, as well as the resting intervals.

The efficiency is computed as:

$$E = S/p$$

The acceleration is a metric of the risk evaluation model's performance and it does not depend on the software or hardware performance. (its value is not influenced by the fact that the algorithm runs as parallel or sequential tasks). The formula is:

$$S = TS/Tp$$

S is computed as a ratio between the time for the fastest sequential algorithm and the necessary time for the same algorithm divided into parallel tasks. Ideally, the acceleration equals the number of processors p, but in practice S is less than p (due to overhead). The case in which S is greater than p reflects an improvement in the memory access by partitioning large data sets between several nodes.

The *cost* is computed as TP (the parallel execution time) multiplied by p (the number of processors).

4. Conclusions

The paper presents some important aspects related to the performance of credit risk evaluation models. three important empirical aspects that need to be taken into account in order to increase the quality of the results are introduced and described: the preliminary analysis of the data sets, the discrimination (the ability of an algorithm to correctly classify the "unviewed" data) and the stability (the optimization of the technique). Consequently, some performance metrics associated to credit risk evaluation applications are explained: the execution time, the overloading, the acceleration and the costs of the application.

Acknowledgements

This article is a result of the project „Doctoral Program and PhD Students in the education research and innovation triangle". This project is co funded by European Social Fund through The Sectorial Operational Programme for Human Resources Development 2007-2013, coordinated by The Bucharest Academy of Economic Studies (project no. 7832, "Doctoral Program and PhD Students in the education research and innovation triangle, DOC-ECI").

References

- [1] M. Golfarelli, D. Maio and D. Maltoni, "On The Error-Reject tradeoff in Biometric Verification Systems", *IEEE PAMI*, Vol. 19, No. 7, pp. 15-25, 1997.
- [2] Z. R. Struzik and A.P.J.M. Siebes, "Wavelet transform based multifractal formalism in outlier detection and localisation for financial time-series", *Physica, A theoretical and statistical physics*, Vol. 2, No. 309, pp. 388-402, 2002.
- [3] Z. R. Struzik, "Wavelet methods in (financial) time-series processing", *Centre for Mathematics and Computer Science (CWI), Amsterdam, The Netherlands, Physica, A theoretical and statistical physics*, Vol. 1, No. 296, pp. 307-319, 2001.
- [4] R. Wilhelm, J. Engblom, A. Ermedahl, N. Holsti, S. Thesing, D. Whalley, G. Bernat, C. Ferdinand, R. Heckmann, T. Mitra, F. Mueller and I. Puaut, "The Worst-Case Execution Time Problem - Overview of Methods and Survey of Tools", *ACM Transactions on Embedded Computing Systems*, Vol. 5, No. 2, pp. 120-140, 2007.
- [5] F. Stappert and P. Altenbernd, "Complete worst-case execution time analysis of straight-line hard real-time programs", *Journal of Systems Architecture*, Vol. 46, No. 4, pp. 339-355, 2000.
- [6] A. R. Henderson, "The bootstrap: A technique for data-driven statistics. Using computer-intensive analyses to explore experimental data", *Clinica Chimica Acta*, Vol. 359, No. 1-2, pp. 1-26, 2005.

Author



Laura POPESCU, PhD Candidate, University of Economics, Bucharest, Romania, Developer, BRD Groupe Societe Generale, Bucharest, Romania. My name is Popescu Laura and I am a PhD Candidate in the second year at The Academy of Economic Studies, Bucharest. I graduated The Faculty of Cybernetics and Informatics in Economy in 2008. My research theme is Artificial Intelligence Methods Applied in Credit Risk Evaluation and I am coordinated by PhD prof. Constanta Bodea. At the same time, I am working as a full-time developer for BRD Groupe Societe Generale. I am interested in topics related to credit risk evaluation, artificial intelligence and financial applications.